Creating Synthetic Database of Poultry and Livestock Operations in Support of Infectious Disease Control Strategies

By Mark Bruhn (RTI)*, Jamie Cajka (RTI),
Dr. Gary Smith (University of Pennsylvania), Ross Curry (RTI),
Seth Dunipace (University of Pennsylvania), William Wheaton (RTI),
Philip Cooley (RTI), Diane Wagener (RTI)

ABSTRACT

Controlling the incidence and propagation of animal-borne infectious disease, such as avian influenza and foot-and-mouth disease, is important to the agriculture industry and public health officials because of the potentially devastating economic consequences and the possibility of transmission to humans. The lack of national geographic information systems (GIS) layers of animal farm locations is hampering efforts to develop control strategies. RTI International collaborated with the University of Pennsylvania through the Models of Infectious Disease Agent Study to generate nationwide spatial layers for several types of livestock and poultry operations that approximate their true sizes and locations. Developing these layers involved generating and placing appropriate numbers of farms within each county based on predictive location surfaces. The number, types, and sizes of the generated farms were derived from the U.S. Census of Agriculture's county totals. These GIS layers allow disease modelers to more realistically simulate outbreaks and compare different control strategies.

BIOGRAPHY

Mr. Bruhn has 18 years of experience as a GIS analyst and applications developer. He has researched and developed software and data solutions to environmental, social, and health challenges. He is developing realistic synthetic poultry and livestock farms to aid infectious disease researchers in their modeling of control strategies.

INTRODUCTION

Good models for the transmission dynamics of infections of veterinary interest, such as avian influenza or foot-and-mouth disease, are created to address specific questions, and most modelers would agree that it is wasteful to create complicated models if simple models will provide an answer that is just as useful. However, modelers who work with animal diseases are increasingly being asked to examine and compare disease control strategies that have an explicit spatial component (e.g., contiguous culling programs or ring vaccination), and they need to respond with models that take into account the spatial relationships between susceptible entities (e.g., farms, ranches, poultry houses). Such stochastic, spatial, and agent-based models are being used with increasing frequency to inform policy decisions—the British models of foot-and-mouth disease are probably the best example—but they all require known farms locations and how many and what type of animals are on each farm.

* Mark Bruhn, 3040 Cornwallis Road Durham, NC 27709; fax: 919-541-8830; mbruhn@rti.org.

In many European countries, these data are available to researchers; however, this is not unconditionally the case in the United States. Only a handful of states have spatial layers of their poultry, swine, and cattle operations. To complicate matters, even when such spatial data exist, their use by disease modelers is severely restricted by producer concerns about data confidentiality. In short, researchers in the United States do not have the kind of access to spatial farm data enjoyed by many of their European colleagues.

While the proposed National Animal Identification System (NAIS) promises to track animals and animal operations, any premises-related information collected under this system will be kept in one or another of five privately maintained databases, and access to this specific information will be provided to state and federal animal health officials only for the purpose of tracking an outbreak. The information stored in the NAIS is exempt from the Freedom of Information Act and will not be shared between inter-governmental agencies. Furthermore, the NAIS program is voluntary with about 20% farmer participation at the moment. Some states, such as Arizona (May 2007, Bill Number S1428), have either already passed or are proposing to present state bills to prohibit mandatory participation in the NAIS.

Attempts to use information that is already in the public domain, such as the addresses of the participants in animal disease eradication campaigns (e.g., bovine tuberculosis or brucellosis) all flounder because of inaccuracies in the lists and because of difficulties inherent in converting rural route addresses to geocodable locations. U.S. Department of Agriculture (USDA) researchers have simulated farm and animal populations in North Carolina by using six spatial constraints, but this effort has only achieved modest success.[1, 2] Other researchers have used satellite imagery to create algorithms for identifying farms of each specific type, but, to date, such methods have created a large number of false positive identifications.

PURPOSE

To create the spatial data required by the models, the geographic information systems (GIS) group at RTI International (RTI) and researchers at the University of Pennsylvania (UPenn) undertook the task to create realistic locations of animal operations from publicly available data. This paper primarily focuses on the first version of synthesized poultry operations created for avian influenza models. Although poultry operations were the first animal operations generated, RTI is currently synthesizing nationwide livestock layers for other types of animal disease control models.

METHOD

A nationwide data layer of realistically located poultry farms will possess the following characteristics: the number of operations by poultry type and size will be realistic, and their locations will be in areas where actual operations of that size and type have a high probability of occurring. The number and type of operations were extracted from the 2002 U.S. Census of Agriculture, which is published by the USDA's National Agriculture Statistics Service (NASS). The probable locations were determined by creating a probability surface from publicly available raster and vector datasets.

The U.S. Census of Agriculture's Web site, which can be accessed at http://www.agcensus.usda.gov/Publications/2002/index.asp, allows users to specify the geographic unit of extraction and table desired. For this project, RTI specified county level summaries for all states (excluding Alaska and Hawaii). We downloaded “Poultry—Inventory and Sales: 2002 and 1997” (Figure 1) from the NASS Web site to obtain counts of poultry operations by type. Similar tables also exist for livestock, such as cattle, hogs, sheep, and goats.

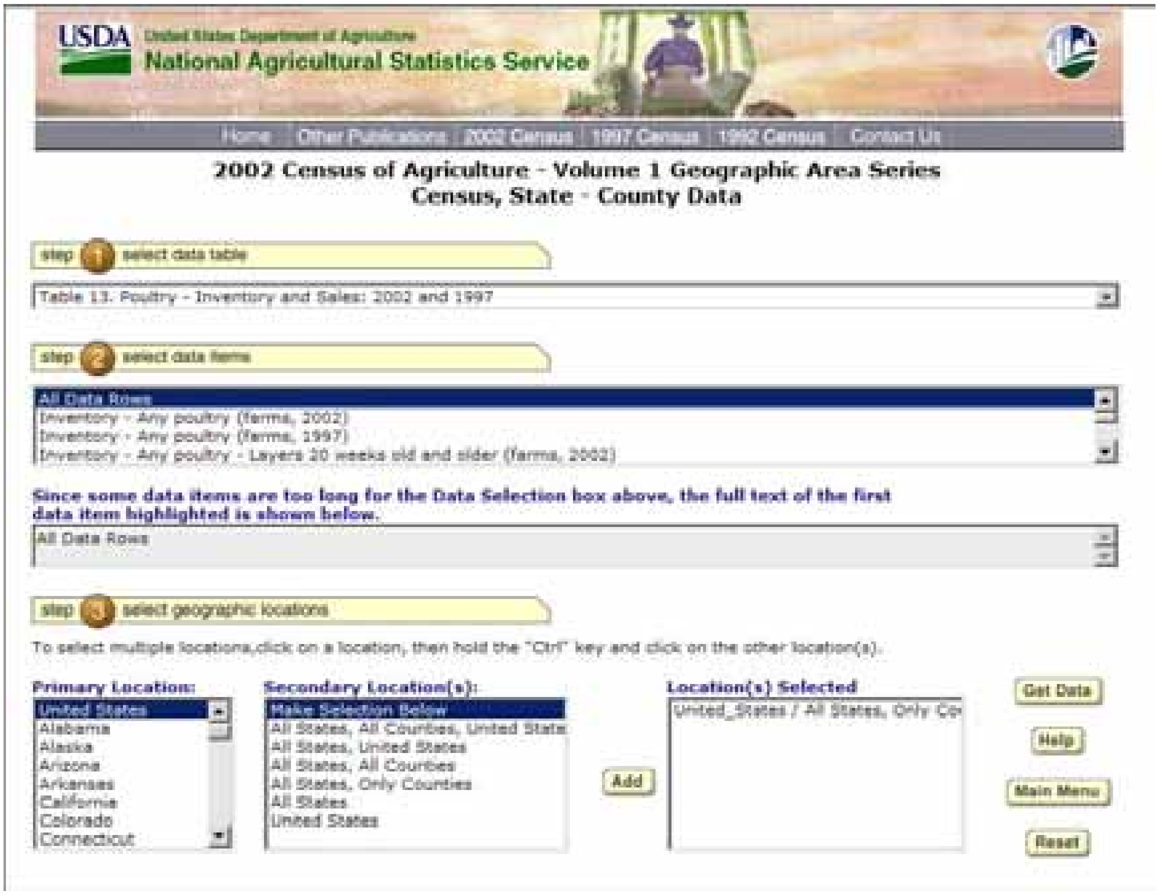


Figure 1. NASS Web site for extracting farm numbers by county.

These data were imported into a Microsoft Access database where summaries and cross-tabulations were performed to create a single record for each county in the United States. Each record contained all of the poultry farm counts for that county. It should be noted that the U.S. Census of Agriculture reports the information in terms of number of farms of a certain animal type. For broilers and laying hens, the farm numbers are further broken out into the number of farms having a certain number of birds (expressed as a range) of that type on or sold by the farm. For example, it reports the number of farms in a given county that have between 1–49 layers 20 weeks or older. By reporting the number of animals by each animal type, the U.S. Census of Agriculture does not specify to what degree the farms may have multiple animal types. If the total number of farms in each category is added together, the sum will be greater than the total number of poultry and livestock farms in the county. This is because some farms raise more than one type of agricultural animal. In the case of poultry, a single farm could produce broilers, layers, and pullets, but they are counted separately in all three categories. The U.S. Census of Agriculture reports the total number of farms engaged in the broad category of raising livestock and poultry, so it is possible to create and place the correct number of operations raising animals. It is not possible to correctly define the degree of mixing of animal subtypes that occurs.

The second data gathering and processing phase was the assemblage of spatial data layers that affect poultry operation locations. Using the research team's past experience, findings from our literature review, and the results of testing against a sample of local and national actual poultry "truth" datasets, we generated the hypotheses about the factors that affect the locations of commercial (particularly large) poultry farms in the conterminous United States (Table 1). Specifically, we did not try to ascertain the factors that determine the location of whole farms, but just the usually large (e.g., 40' x 500') poultry houses on many of these farms because these houses are more spatially constrained than the overall farm. Within a poultry farm, the birds usually reside in these large poultry houses, so we believe focusing on them is also appropriate for using this data as input to disease models.

We used our own experiences to develop the assumptions such that the poultry houses would not be built on water; in the middle of cities; on public lands (being commercial enterprises); on steep slopes, where the building of large buildings would be impractical; in forests or other incompatible land uses, where some other land use was known to exist; and not too close to another poultry farm. The literature review led us to develop some hypotheses about the proximity of poultry farms/houses to accessible roads; sensitive areas, such as parks, schools, and churches; and poultry support services, such as slaughter houses and feed mills. We used a sampling of actual poultry locations from the Dun & Bradstreet (D&B) Business Database and more comprehensive poultry locations for a few counties to test and modify the hypotheses above. The modifications included calibrating the proximities to various inputs (e.g., roads, sensitive areas) to our samples of actual poultry locations.

The comprehensive "truth" data for Lancaster County, PA, was available to researchers at UPenn's School of Veterinary Medicine. This truth data was a set of known poultry operations within Lancaster County, PA, with location, poultry type, and bird number attributes for most of the operations. We also obtained fairly comprehensive locations of poultry farms for the State of Delaware from the Delaware Environmental Navigator, which can be accessed at http://www.nav.dnrec.delaware.gov/dnreceis. We confirmed the locations of the farms by using aerial imagery and moved the points for each farm to the center of the poultry houses on the farm. Based on UPenn researchers' knowledge, and comparisons with the national D&B business data for the State of Delaware, we reasoned that the vast majority of Delaware farms were broiler farms and estimated the number of birds for each farm based on the size and number of poultry houses on each farm. These data were useful for confirming the proximities of the poultry houses to roads. We also had some comprehensive sample data for smaller areas in a handful of other states, but because of time constraints, we were unable to process those sufficiently to use for this version of the dataset.

Table 1. Factors That Are Believed to Affect the Locations of Poultry Houses in the Conterminous United States[3–6]

Areas where poultry houses would not exist include the following:
▪ On incompatible land cover (e.g., on waterbodies, in forests) ▪ On public lands (e.g., military bases, and national, state, and local parks) ▪ On steep slopes where the building of large poultry houses (or any building) is impractical ▪ In highly urbanized areas (e.g., city downtowns) ▪ Where other development already exists (e.g., residential areas, industrial parks, airports, locations of existing businesses) ▪ On other existing land uses (e.g., cultivated crop land).
Areas where poultry houses would be unlikely to exist include the following:
▪ Within municipal boundaries or urbanized areas ▪ Near sensitive areas (e.g., local parks, churches, commercial districts) ▪ Too far from poultry suppliers and service suppliers ▪ Too far from markets and processing plants ▪ Too far from roads accessible by large trucks ▪ Within a minimum distance from one another (50 m).
Areas where poultry houses would be likely to exist include the following:
▪ Within rural areas designated for agricultural use ▪ Near feed suppliers, hatcheries, and poultry service providers ▪ Near poultry processing plants/slaughterhouses ▪ Next to roads accessible by large delivery trucks ▪ Flat areas (<15% slope).

To simulate these factors, we used the datasets listed in Table 2.

Table 2. Site Location Factor Data Layers

Data Layer	Source
Slope	Derived from the National Elevation Dataset
Land cover (including forests and crop lands)	National Land Cover Dataset (NLCD 2001)
Wetlands	National Wetlands Inventory and NLCD 2001
Federal (public) lands	Environmental Systems Research Institute (ESRI) ArcGIS version 9.2 data disks
State and local parks	ESRI ArcGIS version 9.2 data disks
National and state roads	ESRI ArcGIS Business Analyst streetmap (TeleAtlas 2006)
Residential roads	Estimated from ESRI Business Analyst streetmap (TeleAtlas 2006)
Waterbodies	National Hydrography Dataset (medium resolution)
Airports and railroads	ESRI ArcGIS version 9.2 data disks
Poultry support businesses	ESRI ArcGIS Business Analyst (InfoUSA 2006)
Non-agriculture businesses	ESRI ArcGIS Business Analyst (InfoUSA 2006)
Municipalities and urbanized areas	ESRI ArcGIS version 9.2 data disks and U.S. Census Bureau
Sensitive areas (e.g., churches, schools)	ESRI ArcGIS version 9.2 data disks (including Geographic Names Information Service [GNIS] names)

After the initial data gathering was complete, the next step was to project all of the layers into the same USA Contiguous Albers Equal Area Conic U.S. Geological Survey (USGS) projection. All nationwide layers were separated into individual states, which included an extra 1 km buffer around each state. Before rasterizing, the vector shape files were classified to represent only the feature categories that would receive different probabilities. For example, three buffers of increasing distance from roads were assigned unique values. These unique values would later be converted into probabilities that a poultry operation could be located within that class. The resulting state vector layers were rasterized using a 30 m cell size resolution, which matched the NLCD raster layer.

Before we could create individual probability surfaces (in preparation for creating a final composite probability surface), it was necessary to pre-process some of the input data to generate the desired location factor layers. For example, the slope layer was created from the National Elevation Dataset (NED), and then classified into the following three ranges: too steep (slope > 45 degrees), steep (slopes from 15 to 45 degrees), and not steep (slope ≤15 degrees). Existing development (e.g., residential neighborhoods) was simulated by buffering short (less than ½ km) local road segments from a detailed road dataset. Proximity rings were generated at various distances from some input datasets so varying probabilities could be assigned. These rings were created for roads, poultry support service businesses, and waterbodies. The probabilities assigned to each ring were based on the distributions of sample truth locations that fell within them.

The sample truth locations were also used to provide evidence for confirming or refuting our "obvious" assumptions. In particular, it was found that there were some rare cases where our sample truth poultry operations were found on some public lands (e.g., Indian reservations, national forests). We also found a small percentage (<4%) of truth farms falling within the municipal boundaries that we were using. We accommodated these findings by allowing for a small probability that the farms could occur within outer city areas and within public land areas where truth farms had been found.

Individual probabilities of the input layers were then assigned. As previously mentioned, some geographic layers simply precluded the possibility of farms being located within their boundaries. In other cases, there was increasing or decreasing probability of a farm being located near a geographic feature (e.g., road, hatchery, urban area) depending on its nature. Truth data were used to create counts of farms that fell within the various zones, and probabilities were assigned based on these counts.

Once individual raster layers were created with their respective probabilities, they were used in the Single Output Map Algebra tool found in the Spatial Analyst Toolbox of ESRI's ArcGIS version 9.2 software. This tool used all of the individual cell values at any given location and multiplied them together to get a final probability surface (Figure 2). The consequence of multiplying the individual probabilities is that a zero probability in any one of the component layers acted as a filter and excluded those areas from placement consideration. This technique also allowed us to retain areas that might have a low probability for one or two components (e.g., distance from support services) but was optimal in all other layers. This technique also has the advantage over other methods, such as ranking zones within a surface and then simply adding them together, in that any relative differences that might exist between factors do not have to be known or calculated.

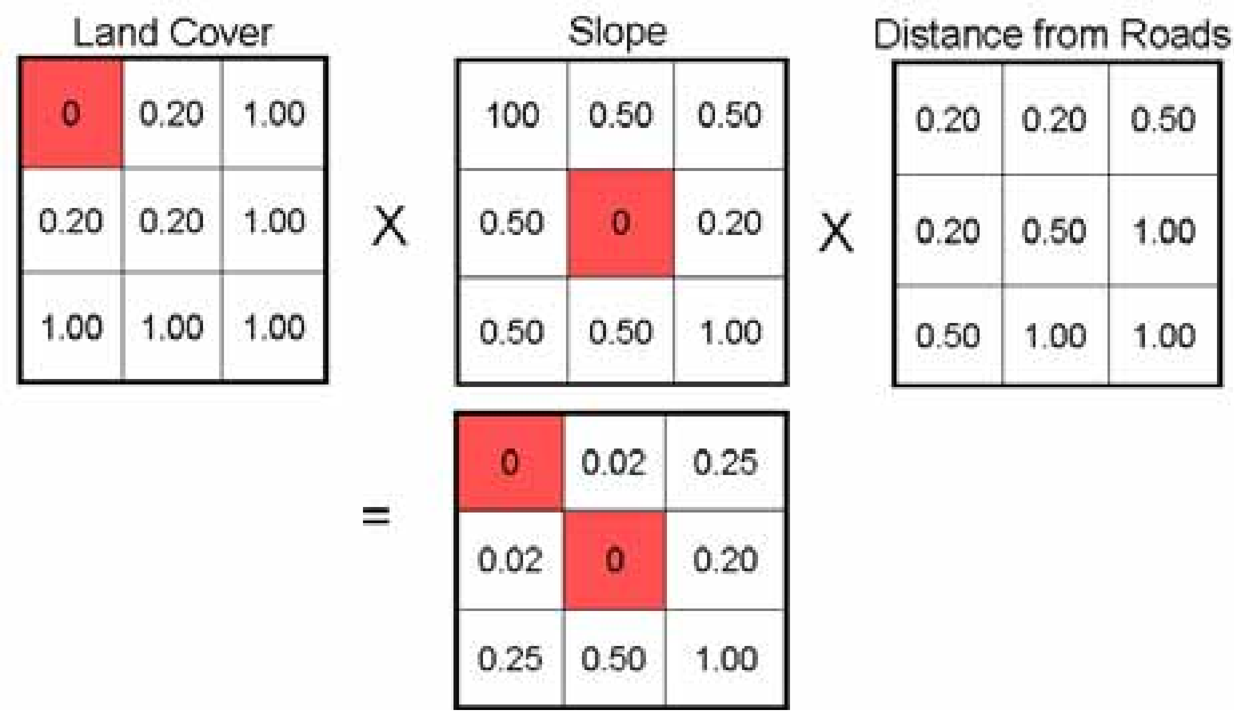


Figure 2. Multiplication of individual probability surfaces.

After determining individual layer probabilities, in accordance with our poultry operation location factor hypotheses and data gathering as described in the data section above, we generated a statewide probability surface for each input data layer. These input data layers were then integrated into a single probability surface against which farms were randomly placed (Figure 3). A single probability surface was used for all poultry types and sizes due to time constraints. If it is determined that further specialization is necessary for the models, we will create a separate probability surface for each animal type and size class that have distinctly different location probabilities.

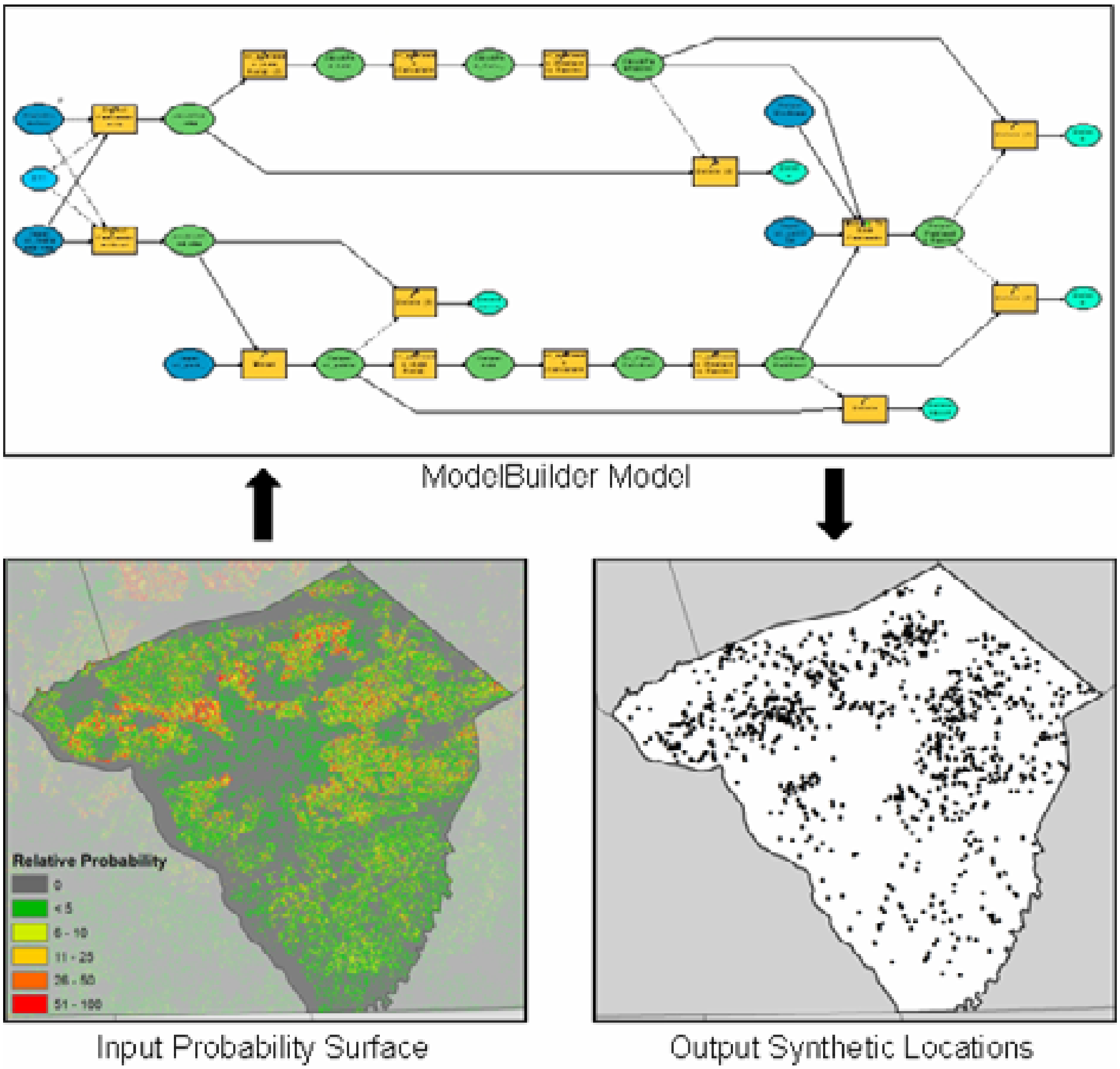


Figure 3. Generating random farm locations using a composite probability surface.

We used ModelBuilder within ArcGIS for pre-processing data, assigning probabilities, and creating probability surfaces and random points. The random point creation process flow included the following major steps:

1. Extracting a polygon(s) for a single county.
2. Clipping a state probability surface with the county polygon(s) to create a county probability surface.
3. Placing 100,000 random points over the county probability surface without regard to probability surface values. (This process could also be performed a point-at-a-time instead of with the batch process we used.)

4. Generating a random number between 1 and the highest probability value present on the county probability surface for each of the 100,000 random points.
5. Extracting the probability surface value for the cell that each point overlaps and inserting it into the point's attribute table for each of the 100,000 points.
6. Removing those points whose random value is greater than the extracted surface probability value.
7. Randomly selecting, from the points that remain, the number of points needed to represent the total number of poultry operations for the given county and deleting the remainder.
8. Attributing each poultry operation point with a poultry type and size (i.e., number of birds) based on the specific poultry type and size farm counts. If it is necessary for the models, we will assign the operation types and sizes in a different manner that is based on specific probability layers.

This process was run from a Visual Basic for Applications (VBA) program that used counts from a table in a Microsoft Access database and then called a model to create the random farm locations. The farm type attribution took place within a module of the VBA project.

RESULTS

The result of this work was a poultry-operation shape file for each county for all of the poultry operations listed in the 2002 U.S. Census of Agriculture in the conterminous United States. While the number of operations and county-level geographic distribution was based on the U.S. Census of Agriculture's figures, the sub-county spatial distribution and composition (number and type of birds) were functions of the probability surface and attribution process.

To test the effectiveness of using a complex set of individual probability surfaces, RTI generated random farm locations in three Pennsylvania counties. Philadelphia County was selected as an example of an urban area, Sullivan County was selected as an example of a non-agricultural and rural location, and Lancaster County was selected as an example of an agricultural and rural area. Two additional sets of farm locations were generated for each county. One set was completely random within the county boundary, and the second set was placed using a simple set of obvious constraints. Figure 4A shows these different placement methods, and Figure 4B shows the random placement of farms using these different options.

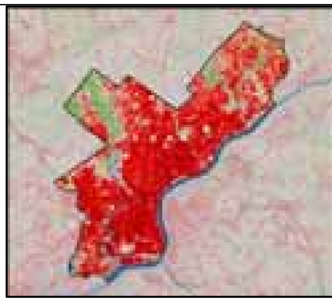
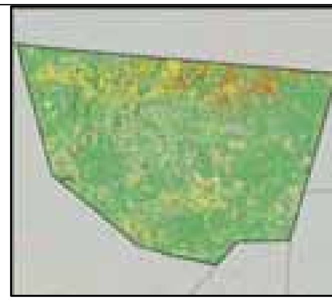
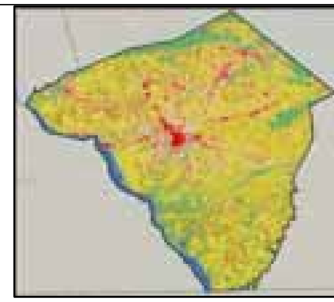
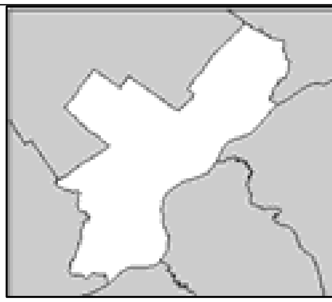
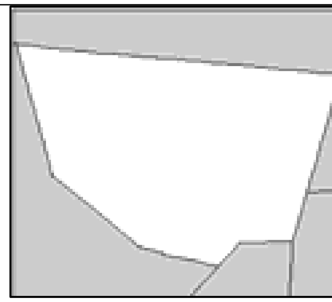
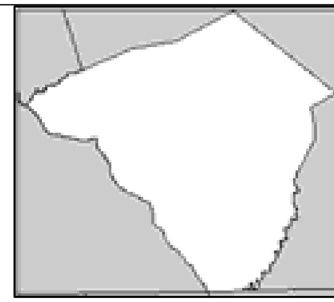
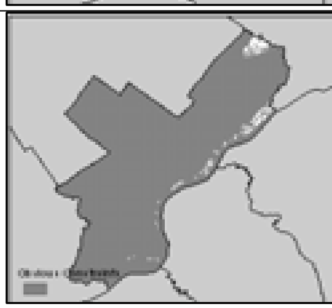
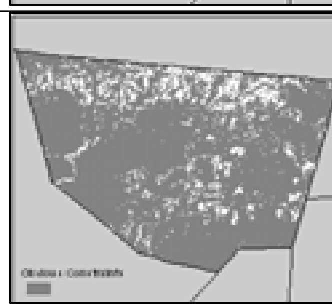
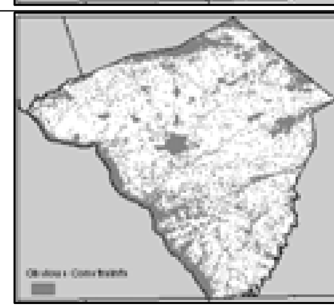
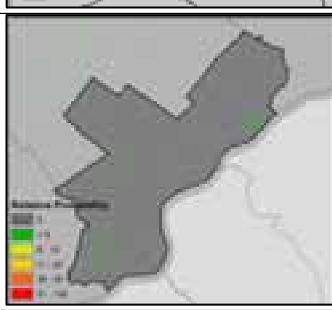
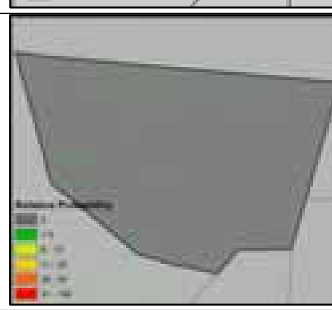
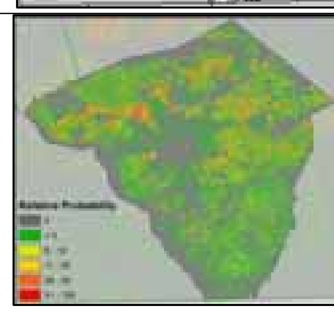
Placement Options	County Types		
	Urban (Philadelphia County, PA)	Rural- Non-agricultural (Sullivan County, PA)	Rural- Agricultural (Lancaster County, PA)
Background Land cover Red = Urban Green = Forest Yellow = Agricultural land			
Across whole county			
Within obvious placement constraints (e.g., forests, urban)			
Using a probability surface (our method)			

Figure 4A. Comparison of our placement method to other less complex methods for different county types: Placement options.

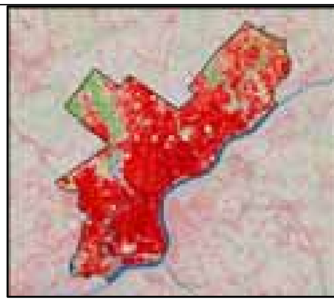
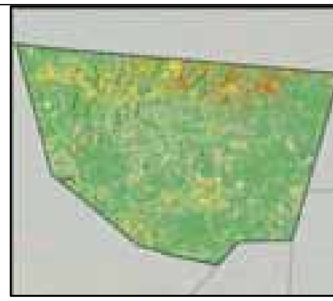
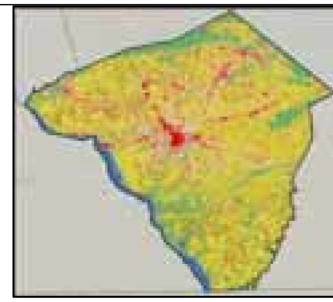
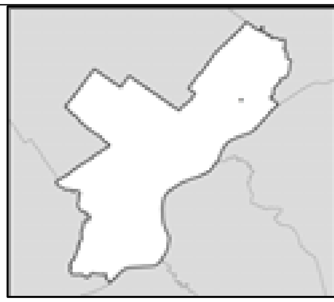
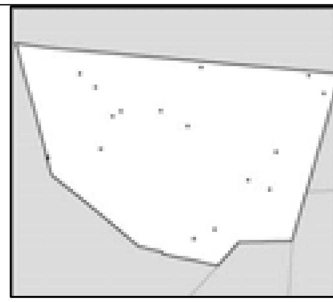
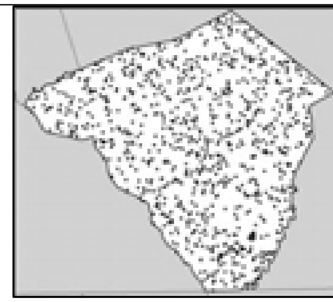
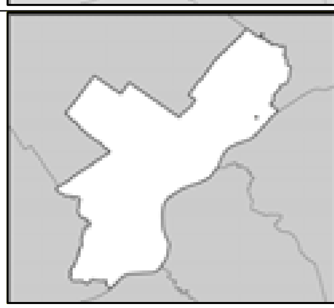
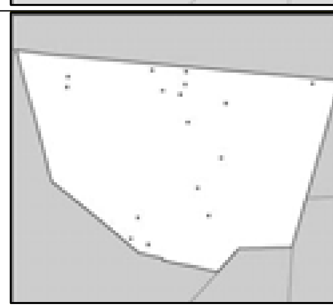
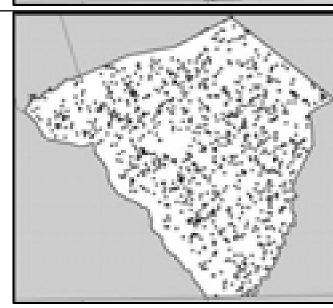
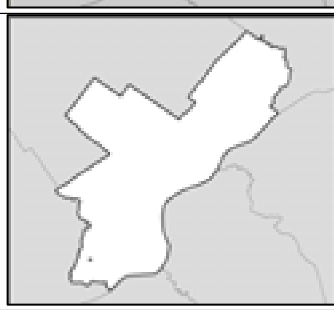
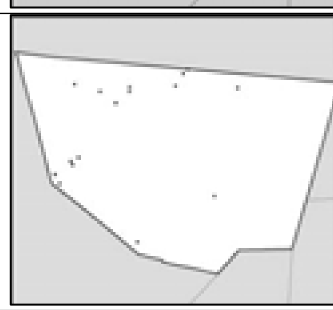
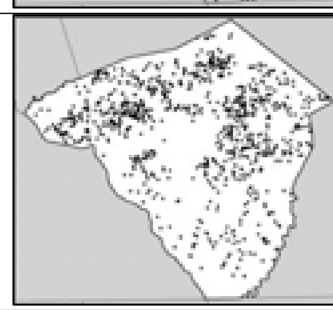
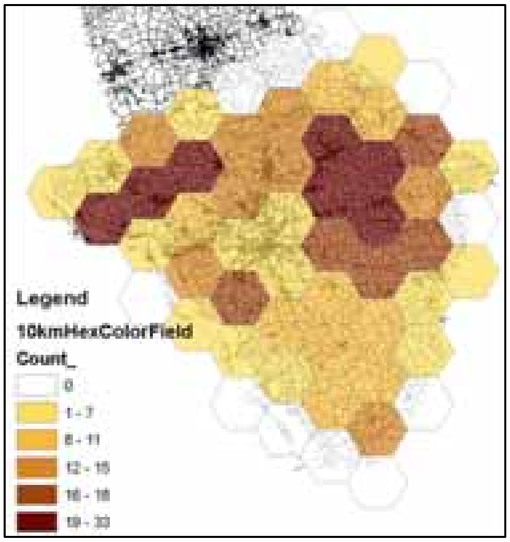
Placement Options	County Types		
	Urban (Philadelphia County, PA)	Rural – Non-agricultural (Sullivan County, PA)	Rural – Agricultural (Lancaster County, PA)
Background Land cover Red = Urban Green = Forest Yellow = Agricultural land			
Random placement across whole county			
Random placement within obvious placement constraints (e.g., forests, urban)			
Random placement using a probability surface (our method)			

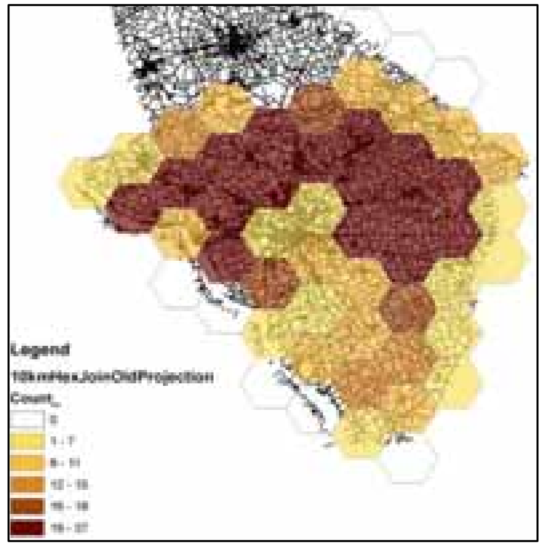
Figure 4B. Comparison of our placement method to other less complex methods for different county types: Randomly placed farm points.

Visual inspection of the RTI Team's method indicated that it created a more focused distribution than other less complex methods, especially for counties with many agricultural areas, where poultry farms are more prevalent. In addition to being more focused than alternative methods, it was also believed that the spatial distribution of the farms placed on the probability surface was similar to the distribution of actual poultry farms. This belief was based on comparisons with the distributions of sample truth data.

To confirm the similarity of the distributions, the synthesized poultry operation data were sent to researchers Dr. Gary Smith and Mr. Seth Dunipace at UPenn's School of Veterinary Medicine. They evaluated the synthesized locations against the more comprehensive Lancaster Truth Dataset. Figure 5 shows the actual and synthesized poultry operations summarized to a 10 km hexagonal grid overlaying Lancaster County. Actual poultry operation locations are not shown due to confidentiality restrictions.



Truth Locations



Synthesized Locations

Figure 5. Lancaster County, PA, poultry operation locations summarized to 10 km hexagonal grids.

The UPenn researchers found that overall the distribution of synthesized poultry operations was generally in agreement with the actual poultry distribution for the Lancaster County pilot area. The size and type of synthesized operations did not match the actual ones because these attributes were randomly assigned within the U.S. Census of Agriculture's county-level count constraints. It is unknown whether this will make a significant difference in the model outcomes, and further testing is still required in this matter. If the differences are found to be significant, we can use separate probability surfaces for each significantly different size and poultry type combination group.

CONCLUSION

It is possible to generate rudimentary poultry-operation locations nationwide using U.S. Census of Agriculture data and GIS. Using obvious constraints such as excluding waterbodies, forests, and urban areas from placement consideration will result in a more realistic distribution of poultry operations for non-agricultural counties over unconstrained placement across a whole county. Going one step further, by incorporating additional layers that specifically pertain to poultry operations, such as poultry support services, and assigning probabilities that vary with distance, this adds a significant level of refinement that results in a distribution much closer to that of known operations. This is especially true in the agricultural counties, where it is believed that a majority of poultry operations occur and the concern over the control of animal diseases is the greatest.

While the locations and attribution are not perfect, they are a reasonable surrogate for actual farm locations. We believe that more probability surfaces customized for poultry subtypes (e.g., broilers, layers, pullets) and sizes (e.g., small, medium, large) will result in more realistic distributions and attribution. This process is also being developed for livestock. Synthesized cattle (beef and dairy) operations are being developed, with sheep, goats, and hogs to follow. It is also possible to depict mixed farms, where an inventory of all livestock types will be simulated. Feedback from disease model users and developers will guide the data development process. Knowing what aspects of the data affect the models the most, will allow data developers to know where to focus their future efforts.

REFERENCES

1. Miller, R., K. Geter, B. Corso, K. Forde-Folle, D. Neufeld, J. Freier, and J. Russell. 2007. Farm location and animal population simulator: A system for estimating farm and animal populations. Poster presented at the GISVet 2007 Conference, Copenhagen, Denmark.

2. Geter, K. 2006. Farm animal demographics simulator aids in disease modeling. NAHSS Outlook, Quarter Two.

3. Berry, J. 2007. Factors involved in site selection for new and modified poultry facilities. Oklahoma Cooperative Extension Fact Sheet F-8213. May.

4. Goan, C. Site Selection Factors for New Poultry Facilities. The University of Tennessee Agricultural Extension Service, SP 592.

5. Sambidi, P.R., R.W. Harrison, and A.J. Farr. 2004. A Conjoint analysis of site selection for the U.S. broiler industry: Implications for Louisiana. Louisiana State University Agricultural Center Bulletin Number 882. August.

6. Sambidi, P.R. 2003. Factors affecting plant location decisions of U.S. broiler executives. Master’s Thesis, Louisiana State University. May.